数理解析研究所講究録
第 519 巻 1984 年 99-113

Limited Processor Sharing Queues

工学院大 山崎 源治 (Genji Yamazaki)
筑波大 逆瀬川浩孝 (Hirotaka Sakasegawa)

1. はじめに

外部から到着するジョブを処理する場合，1人のサーバを配するか，その $1/m$ の能力を持つ m 人 ($m \geq 2$) のサーバを配すべきか，という問題は，待ち行列論で古くからよく議論されてきた問題の1つである。もちろん，m-サーバシステムでの"待ち"の解析は非常に難しいため，単一サーバとの定量的比較ができるのは，ごく限られた場合のみである。しかしながら，単一サーバシステムでは，システム内に1つでもジョブが存在するときは常にそのサービス能力をフルに使うが，m-サーバシステムでは，システム内に m 以上のジョブが存在するときのみそのサービス能力をフルに使うが，ジョブ数が m 未満のときはその1部のみを使うことから，直観的には，"待ち"(ジョブの列待ち時間ではなく，その系内滞留時間などシステム全体に関するもの)からみて，単一サーバシステムの方が有効であると考えられる。

この観点から，Stidham[3]は，単一サーバシステムとm-サーバシステムの系内人数分布の比較を試み（両システムの利用率は等しいという条件で），ジョブのサービス時間がアーラン分布に従う場合，単一サーバの系内人数はm-サーバのそれよりも確率的に小さい（単一サーバの有効性）を示した。この結果はより一般的なサービス時間分布のもとでも成立しそうであるが，サービス時間の変動係数が1以上のときには反例すら提出されていて（Brumelle[1]），上述の結果が成立するサービス時間の分布のクラスは未だ明らかではないように思われる。

この単一サーバの有効性は，m-サーバシステムでは，システム内ジョブ数がm未満のときいわゆる“手空き”のサーバが存在していることに起因しているが，m-サーバシステムにおいて“ジョブが1つでもシステム内に存在するときは手空きのサーバはいない”という規律のもとで，いいかえると“手空きのサーバは他へ手伝いに行く”という規律のもとでジョブを処理する場合も単一サーバの有効性が保たれるか否か，という問題が当然生じてくる。

より具体的に述べると，単一サーバシステムと次のように作動する2-サーバシステムを比較したとき，単一サーバの有効性が保たれるか，否か，の問題である。システムが空の

状態でジョブが到着したとき，2人のサーバでそのジョブを処理する（そのときのサービスレイトは1人のときの2倍），その処理中に他のジョブが到着したなら，1人のサーバは直ちにその新しいジョブの処理を開始する。2つのジョブの処理中に新しいジョブが到着したなら，それらのジョブは先着順（FCFS）で待つ。もちろん，新しいジョブの到着前に2つのジョブのどちらかの処理が終了したなら，直ちに他のジョブの処理を2人のサーバで行う。

ここで，1つのジョブを2人で処理する際のサービスレイトが1人のときの2倍，という設定はかなり理想的であるかも知れないが，いわゆる“手空きのサーバは他へ手伝いに行く”ということは，実際にも数多く見られる。また，上述の2-サーバシステムのように作動する m-サーバシステムと単一サーバシステムの比較の問題は，計算機のCPUの多重度という観点でより自然に解釈できる。すなわち，あるCPUでジョブを1つずつFCFSで処理する場合が単一サーバシステムに相当し，mジョブまではFCFSでタイム・シェアにより処理し，それをこえたとき待ちが生じる場合がm-サーバシステムに相当しているわけである。

上述の問題に対して，1つの答えを与えることが本稿の目的である。

2. モデル，記号

本稿では，次のように作動する待ち行列システムを扱う。システム内のスペースを適当に分割して，前から順にpostions 1，2，… と番号づけする（1つのpositionには，1ジョブのみ入ることができる）。システム内にnジョブいるとき，

(i) 到着したジョブはposition $(n+1)$へ入る（FCFS），

(ii) 正の整数mをあらかじめ固定したとき，サーバ（1人のみ）はそのサービス能力をposition jに，

$\alpha_j(n)$ (>0)，$\sum_{j=1}^{n}\alpha_j(n)=1$　for $n \leq m$，

$\alpha_j(n)$ $(>0, j=1,2,\cdots,m)$，$\alpha_j(n)=0$ $(j=m+1,\cdots,n)$

$\sum_{j=1}^{m}\alpha_j(n)=1$　for $n>m$，

なる比率α. だけふり向ける，

(iii) position jのジョブの処理が終了したなら，positions $j+1, j+2, \cdots, n$にいたジョブは，順にpositions $j, j+1, \cdots, n-1$へ移る。

(ii)をより具体的に解釈すると，上述のようにシステム内を分割したとき，ある固定したmに対して，positions $1, 2, \cdots, m$はservice positionsとなり，positions $m+1, m+2, \cdots$がwaiting positionsとなる（Fig. 1）。このため，以下では上述のシステムをLimited Processor Sharing Queue（LIPSと略す）と呼び，特に$m=1$のときを1-LIPS，$m(\geq 2)$をm-LIPS

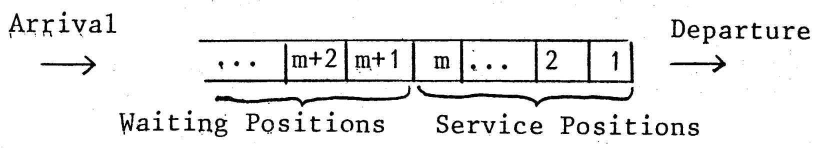

Fig. 1. Limited Processor Sharing Queue

と呼ぶことにする。

1-LIPS, m-LIPS に対して，以下では次の記号を用いる。

$L_t(1)$ = 1-LIPSでの時刻 t におけるシステム内ジョブ数,
$L_t(m)$ = m-LIPS 〃 ,
$W_t(1)$ = 1-LIPSでの t における残り仕事量の総和,
$W_t(m)$ = m-LIPS 〃

ジョブの処理時間は，i.i.d. r.v.'s でその r.v を S で表わし，その分布関数を $B(\cdot)$ とする。ここでの分布関数の意味は，システム内のサーバが1つのジョブのみの処理に集中したときの確率である。またジョブのシステムへの到着過程は次の条件を満たすものとする（到着間隔が i.i.d. r.v.s である必要がないことに注意）.

[A-1]：到着過程は処理とは独立で，有限時間区間内に到着するジョブ数は確率1で有限.

一見，LIPS での (ii) の規律は抽象的であるが，α_i を適当に設定すると，次のような通常の待ち行列システムとなる。

例・1　$m=1 \Rightarrow$ FCFS の G/GI/1 queue.

例・2　$n \leq m$ のとき $\alpha_j(n) = 1/n$，$n > m$ のとき $\alpha_j(n) = 1/m$ $(j=1,2,\cdots,m)$，$\alpha_j(n) = 0$ $(j \geq m+1)$ とすると，FCFS の多重度 m のタイムシェア処理の CPU，特に $m=2$ のときは§1 で述べた"手空きのサーバは他へ手伝いに行く"という 2-サーバシステム．

到着間隔（i.i.d. r.v's）の分布，$B(x)$ を明示するときは，ケンドールの記号を用いる。例えば，$M/E_2/m$ (LIPS) により，到着間隔が指数分布，$B(x)$ がフェーズ 2 の m-LIPS を表わす。

3. 基礎的結果

3.1 保存則

ジョブの1-LIPSへの到着時点列を$\{t_i\}_{i=0,1,2,\cdots}$ $(t_0=0)$とする。相続くジョブの処理時間を$S_0, S_1, S_2, \cdots$としたとき，$W_t(1)$を図示するとt_iでS_iだけジャンプし，他では45°のスロープで減る，いわゆるG/GI/1の仮待ち時間過程となる。一方，同じジョブがm-LIPSに入るとすると，$W_t(m)$はt_iでS_iだけジャンプし，他では45°のスロープで減る（$\sum \alpha_j(\cdot)=1$より）ため，$W_t(1)$と全く一致する。すなわちLIPSは"保存則"が成立するシステムである。さらに，$L_t(m)=0$なる事象は$W_t(m)=0$と一致するため，$W_{0-}(1)=W_{0-}(m)$のとき（便宜上，以上では$W_{0-}(1)=W_{0-}(m)=0$とする），次の結果を得る。

定理1

$$W_t(1) \overset{d}{=} W_t(m), \quad Pr(L_t(1)=0)=Pr(L_t(m)=0),$$

ここで，$\overset{d}{=}$は分布の一致を意味する。

3.2 アーラン・サービス

本節では，ジョブの処理時間の分布がフェーズkのアーラン分布（E_k；1つのフェーズのレイトは$k\mu$）である場合の$L_t(1)$と$L_t(m)$の比較を試みる。

いま，ジョブの到着時点列$t_0, t_1, \cdots, t_{n+1}$が与えられたときの時刻$t$ ($t_n \leq t < t_{n+1}$) における1-LIPSとm-LIPSの残りフェーズ数の総和（待っているジョブについては1ジョブにつきk，処理中のジョブに関しては，現在処理中のフェーズを含めた残りフェーズ数）を比較する。げん密な議論は省略し直観的な解釈のみを記す。時刻u ($<t$) に1-LIPSに1つでもジョブが存在するなら，$[u, u+\Delta u)$ ($\Delta u > 0$) にフェーズ数が1つ減る確率は，

$$k\mu\Delta u + o(\Delta u)$$

で与えられる。m-LIPSにおいても，$L_u(m) \geq 1$なら，$[u, u+\Delta u)$でフェーズが1つ減る確率は，

$$\sum_{j=1}^{L_u(m)} \alpha_j(L_u(m))k\mu\Delta u + o(\Delta u) = k\mu\Delta u + o(\Delta u)$$

となり，1-LIPSのそれと一致する。到着時点列を両システムでそろえていることから，両システムに到着する総仕事量（総フェーズ数）の構造は全く一致する。これらの事実と，定理1より両システムでuでシステムが空である確率は等しいことから，到着に関する条件をはずすことにより，次の結果を得る。

補助定理． $B(\cdot) = E_k$ のとき，$N_t(1)$, $N_t(m)$を時刻tにおける1-LIPS, m-LIPS の残りフェーズ数とする。このとき，

$N_k(1) \stackrel{st}{\geq} N_k(m)$.

この結果を用いて，$L_\pi(1)$と$L_\pi(m)$の比較を試みる。システム内にnジョブいるときの1-LIPS，m-LIPSにおける残りフェーズ数の総和を，それぞれ$N(1,n)$，$N(m,n)$で表わす。このとき，起り得るすべての組合せを考えると，$N(1,n)$，$N(m,n)$は次の不等式を満たす整数となることが容易に確かめられる。

$$(n-1)k+1 \leq N(1,n) \leq nk,$$

$$\begin{cases} n \leq N(m,n) \leq nk & \text{for } n \leq m \\ (n-m)k+m \leq N(m,n) \leq nk & \text{for } n > m. \end{cases}$$

この不等式から，残りフェーズ数Nを固定したときの1-LIPSでの可能なジョブ数$n(1)$とm-LIPSでのそれ$n(m)$を求めると次の等号あるいは不等式を満たす整数となる。

$$n(1) = \left[\frac{N+k-1}{k}\right],$$

$$\begin{cases} \left[\frac{N+k-1}{k}\right] \leq n(m) \leq N & \text{for } N \leq m, \\ \left[\frac{N+k-1}{k}\right] \leq n(m) \leq \left[\frac{N+m(k-1)}{k}\right] & \text{for } N > m, \end{cases}$$

ここで任意の実数aに対して$[a]$は，aをこえない最大整数（Gaussの記号）を意味する。すなわち，Nを固定したとき$n(1) \leq n(m)$となり，これと，上述の補助定理から次の結果を得る。

定理2. $B(\cdot)=E_k$ なら,

$$L_\tau(1) \overset{st}{\leq} L_\tau(m) \quad \text{for all } \tau,$$

ここで, $\overset{st}{\leq}$ は確率的大小関係を示す。

この結果は, LIPSでも, 通常の複数サーバシステムと単一サーバシステムを比較したときの単一サーバの有効性([3])が保たれることを意味している。$L_\tau(m-1) \overset{st}{\leq} L_\tau(m)$ も成立しそうであるが, ここで用いた証明法では得ることができなかった。これについては§4で数値的に検討する。また, $L_\tau(1)$ と $L_\tau(m)$ の定量的比較, 及び処理時間の変動係数 C_s が1以上の場合は逆の順序関係となることを, §3.2, §3.3で例示する。

3.2 $M/E_2/2$ (LIPS)

到着間隔が指数分布(レイト λ), ジョブの処理時間の分布がフェーズ2のアーラン分布(1つのフェーズのレイトは 2μ)の2-LIPSを考える。このシステムの定常状態での平均システム内ジョブ数 $L(2)$ を求めるため,

$$F(z) \overset{\Delta}{=} P_0 + P_1 z + P_2 z^2 + \cdots,$$

ここで $P_n = \lim_{\tau\to\infty} P_r(L_\tau(2)=n)$, なる母関数 $F(z)$ を導入する。

この$F(z)$は，P_nに関する平衡方程式から導くことができるが，以下ではＳ２の例２の場合について$F(z)$を導くと次のようになる（その際$\rho=\lambda/\mu<1$，定理１から$P_0=1-\rho$を用いる）。

$$F(z)=\frac{C(z)}{D(z)}(1-\rho),$$

$$C(z)=\frac{\rho^3}{2}z^3-\rho^2(2+\rho)z^2+\frac{\rho}{2}(8+4\rho+\rho^2)z-4(2+\rho),$$

$$D(z)=\rho^3z^3-2\rho^2(3+\rho)z^2+\rho(12+6\rho+\rho^2)z-4(2+\rho).$$

この$F(z)$から，$L(2)=F'(1)$として$L(2)$を求めることができ，それは次式で与えられる。

$$L(2)=\frac{\rho(4-\rho^2)}{4(1-\rho)}.$$

一方，この場合の1-LIPSの平均システム内ジョブ数（定常状態）$L(1)$は，$M/E_2/1$のそれとなるため，既知の結果から，

$$L(1)=\frac{\rho(4-\rho)}{4(1-\rho)}$$

となる。もちろん，$L(2)/L(1)=1+\frac{\rho(1-\rho)}{4-\rho}>1$であるが，この比は$\rho=0.535\cdots$で最大となる。すなわち，この場合の単一サーバの有効性はρに関して単調ではなく，$\rho=0.5$付近で

それが顕著となることを注意する。

3.3 M/H/2(LIPS)

到着間隔が指数分布で，$B(x)$が次の超指数分布のときの$L(2)$も母関数を経て求めることができる。

$$B(x)=\alpha(1-e^{-2\alpha\mu x})+(1-\alpha)(1-e^{-2(1-\alpha)\mu x}).$$

$\nu_2(\cdot)$を前節と同様に§2の例2の場合について計算すると，

$$L(2)=L(1)-\varepsilon \qquad (\varepsilon>0),$$

ここで$L(1)$は1-LIPSの場合の平均系内ジョブ数である。すなわち，この場合($C_s^2>1$)は，単一サーバの有効性は成立せず，2-LIPSの方が有効となる。

4. 数値例

m-LIPSの平均システム内ジョブ数を$L(m)$で表わしたとき，いろいろな場合についての数値実験を試み，$L(m)/L(1)$についての結果を得ているが，それによると$C_s^2<1$のときは，$L(m-1)<L(m)$，$C_s^2>1$のときは$L(m-1)>L(m)$とな

[表・1] $E_5/E_2/m$ (LIPS)

ρ	$L(2)/L(1)$	$L(4)/L(1)$
0.3	1.0149	1.0158
0.5	1.0482	1.0606
0.7	1.0750	1.1257
0.9	1.0518	1.1263

[表・2] $M/H/m$ (LIPS) $C_s^2=3$

ρ	$L(3)/L(1)$	$L(5)/L(1)$
0.3	0.783	
0.5	0.735	0.681
0.7	0.777	0.676
0.9	0.907	0.833

るように思われる。その一例を[表・1]，[表・2]で示す。[表・2]の$C_s^2=3$は，§3.3のHでαを適当に設定している。すなわち，処理時間のバラツキが小さい場合は，単一サーバの方が

有効となるが，バラツキが　きくなると，mを増した方が有効となるようである。

5. おわりに

LIPSにおけるmの増加が待ちにどのような影響を与えるかについて，若干の部分的結果を得ることができた。しかしながら，§2の方法は，フェーズの特性に強く依存していて他の分布のクラスについては適用が難かしく，また$L_\pi(m-1)$と$L_\pi(m)$の比較もできなかった。数値結果によると，$L_\pi(m-1) \overset{st}{\leq} L_\pi(m)$となるようであるが，これは今後の課題である。

また$C_s^2>1$についての特性の解析は，M/H/2(LIPS)のみで，数値実験の結果に頼らざるを得なかったが，最近$B(\cdot)$がNWUであるとき，すなわち$\bar{B}(x)\bar{B}(y) \leq \bar{B}(x+y)$のとき$L_\pi(1) \overset{st}{\leq} L_\pi(m)$となることが証明されたことを付記する(Yamazaki [4]).

$\nu_d(\cdot)$については，数値実験等では§2の例.2の場合についてのみ扱っているが，§3.2の結果は$\nu_d(\cdot)$には依存しないし，また[4]の結果も同様である。ここでは，FCFSのLIPSのみを扱ったが，サーバの能力を各positionに確率的にふり向け，ジョブの各positionへの入るのも確率的に定めた

ときも同様の結果が成立する(Kelley [2]はそれらの確率は等しい場合を扱ったが，ここでは独立に定めてもよい)。

References

1. Brumelle, S. L., "Some inequalities for parallel service queues," Opns. Res., 19 (1971).

2. Kelly, F. P., "Networks of queues," Adv. Appl. Prob., 8 (1976).

3. Stidham, S. Jr., "On the optimality of single-server queueing systems," Opns. Res., 18 (1970).

4. 山崎，"LIPSにおける確率的順序関係"，日本OR学会春季研究発表会(1984，予定).